Improving Entity Linking by Modeling Latent Relations between Mentions

Improving Entity Linking by Modeling Latent Relations between Mentions

Phong Le[1] **and Ivan Titov**[1,2]
[1]University of Edinburgh [2]University of Amsterdam
{ple,ititov}@inf.ed.ac.uk

Abstract

Entity linking involves aligning textual mentions of named entities to their corresponding entries in a knowledge base. Entity linking systems often exploit relations between textual mentions in a document (e.g., coreference) to decide if the linking decisions are compatible. Unlike previous approaches, which relied on supervised systems or heuristics to predict these relations, we treat relations as latent variables in our neural entity-linking model. We induce the relations without any supervision while optimizing the entity-linking system in an end-to-end fashion. Our multi-relational model achieves the best reported scores on the standard benchmark (AIDA-CoNLL) and substantially outperforms its relation-agnostic version. Its training also converges much faster, suggesting that the injected structural bias helps to explain regularities in the training data.

1 Introduction

Named entity linking (NEL) is the task of assigning entity mentions in a text to corresponding entries in a knowledge base (KB). For example, consider Figure 1 where a mention "World Cup" refers to a KB entity FIFA_WORLD_CUP. NEL is often regarded as crucial for natural language understanding and commonly used as preprocessing for tasks such as information extraction (Hoffmann et al., 2011) and question answering (Yih et al., 2015).

Potential assignments of mentions to entities are regulated by semantic and discourse constraints. For example, the second and third occurrences of mention "England" in Figure 1 are coreferent and thus should be assigned to the same entity. Besides coreference, there are many other relations between entities which constrain or favor certain alignment configurations. For example, consider relation participant_in in Figure 1: if "World Cup" is aligned to the entity FIFA_WORLD_CUP then we expect the second "England" to refer to a football team rather than a basketball one.

NEL methods typically consider only coreference, relying either on off-the-shelf systems or some simple heuristics (Lazic et al., 2015), and exploit them in a pipeline fashion, though some (e.g., Cheng and Roth (2013); Ren et al. (2017)) additionally exploit a range of syntactic-semantic relations such as apposition and possessives. Another line of work ignores relations altogether and models the predicted sequence of KB entities as a bag (Globerson et al., 2016; Yamada et al., 2016; Ganea and Hofmann, 2017). Though they are able to capture some degree of coherence (e.g., preference towards entities from the same general domain) and are generally empirically successful, the underlying assumption is too coarse. For example, they would favor assigning all the occurrences of "England" in Figure 1 to the same entity.

We hypothesize that relations useful for NEL can be induced without (or only with little) domain expertise. In order to prove this, we encode relations as latent variables and induce them by optimizing the entity-linking model in an end-to-end fashion. In this way, relations between mentions in documents will be induced in such a way as to be beneficial for NEL. As with other recent approaches to NEL (Yamada et al., 2017; Ganea and Hofmann, 2017), we rely on representation learning and learn embeddings of mentions, contexts and relations. This further reduces the amount of human expertise required to construct the system and, in principle, may make it more portable across languages and domains.

Our multi-relational neural model achieves an

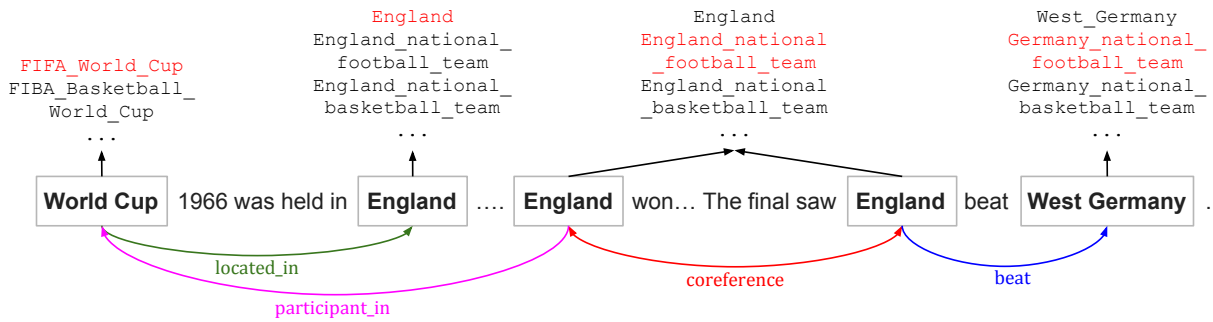

Figure 1: Example for NEL, linking each mention to an entity in a KB (e.g. "World Cup" to FIFA_WORLD_CUP rather than FIBA_BASKETBALL_WORLD_CUP). Note that the first and the second "England" are in different relations to "World Cup".

improvement of 0.85% F1 over the best reported scores on the standard AIDA-CoNLL dataset (Ganea and Hofmann, 2017). Substantial improvements over the relation-agnostic version show that the induced relations are indeed beneficial for NEL. Surprisingly its training also converges much faster: training of the full model requires ten times shorter wall-clock time than what is needed for estimating the simpler relation-agnostic version. This may suggest that the injected structural bias helps to explain regularities in the training data, making the optimization task easier. We qualitatively examine induced relations. Though we do not observe direct counterparts of linguistic relations, we, for example, see that some of the induced relations are closely related to coreference whereas others encode forms of semantic relatedness between the mentions.

2 Background and Related work

2.1 Named entity linking problem

Formally, given a document D containing a list of mentions $m_1, ..., m_n$, an entity linker assigns to each m_i an KB entity e_i or predicts that there is no corresponding entry in the KB (i.e., $e_i = \text{NILL}$).

Because a KB can be very large, it is standard to use an heuristic to choose potential candidates, eliminating options which are highly unlikely. This preprocessing step is called *candidate selection*. The task of a statistical model is thus reduced to choosing the best option among a smaller list of candidates $C_i = (e_{i1}, ..., e_{il_i})$. In what follows, we will discuss two classes of approaches tackling this problem: local and global modeling.

2.2 Local and global models

Local models rely only on local contexts of mentions and completely ignore interdependencies between the linking decisions in the document (these interdependencies are usually referred to as *coherence*). Let c_i be a local context of mention m_i and $\Psi(e_i, c_i)$ be a local score function. A local model then tackles the problem by searching for

$$e_i^* = \arg\max_{e_i \in C_i} \Psi(e_i, c_i) \quad (1)$$

for each $i \in \{1, ..., n\}$ (Bunescu and Paşca, 2006; Lazic et al., 2015; Yamada et al., 2017).

A *global* model, besides using local context within $\Psi(e_i, c_i)$, takes into account entity coherency. It is captured by a coherence score function $\Phi(E, D)$:

$$E^* = \arg\max_{E \in C_1 \times ... \times C_n} \sum_{i=1}^{n} \Psi(e_i, c_i) + \Phi(E, D)$$

where $E = (e_1, ..., e_n)$. The coherence score function, in the simplest form, is a sum over all pairwise scores $\Phi(e_i, e_j, D)$ (Ratinov et al., 2011; Huang et al., 2015; Chisholm and Hachey, 2015; Ganea et al., 2016; Guo and Barbosa, 2016; Globerson et al., 2016; Yamada et al., 2016), resulting in:

$$E^* = \arg\max_{E \in C_1 \times ... \times C_n} \sum_{i=1}^{n} \Psi(e_i, c_i) + \sum_{i \neq j} \Phi(e_i, e_j, D) \quad (2)$$

A disadvantage of global models is that exact decoding (Equation 2) is NP-hard (Wainwright et al., 2008). Ganea and Hofmann (2017) overcome this using loopy belief propagation (LBP),

an approximate inference method based on message passing (Murphy et al., 1999). Globerson et al. (2016) propose a *star model* which approximates the decoding problem in Equation 2 by approximately decomposing it into n decoding problems, one per each e_i.

2.3 Related work

Our work focuses on modeling pairwise score functions Φ and is related to previous approaches in the two following aspects.

Relations between mentions

A relation widely used by NEL systems is *coreference*: two mentions are coreferent if they refer to the same entity. Though, as we discussed in Section 1, other linguistic relations constrain entity assignments, only a few approaches (e.g., Cheng and Roth (2013); Ren et al. (2017)), exploit any relations other than coreference. We believe that the reason for this is that predicting and selecting relevant (often semantic) relations is in itself a challenging problem.

In Cheng and Roth (2013), relations between mentions are extracted using a labor-intensive approach, requiring a set of hand-crafted rules and a KB containing relations between entities. This approach is difficult to generalize to languages and domains which do not have such KBs or the settings where no experts are available to design the rules. We, in contrast, focus on automating the process using representation learning.

Most of these methods relied on relations predicted by external tools, usually a coreference system. One notable exception is Durrett and Klein (2014): they use a joint model of entity linking and coreference resolution. Nevertheless their coreference component is still supervised, whereas our relations are latent even at training time.

Representation learning

How can we define local score functions Ψ and pairwise score functions Φ? Previous approaches employ a wide spectrum of techniques.

At one extreme, extensive feature engineering was used to define useful features. For example, Ratinov et al. (2011) use cosine similarities between Wikipedia titles and local contexts as a feature when computing the local scores. For pairwise scores they exploit information about links between Wikipedia pages.

At the other extreme, feature engineering is almost completely replaced by representation learning. These approaches rely on pretrained embeddings of words (Mikolov et al., 2013; Pennington et al., 2014) and entities (He et al., 2013; Yamada et al., 2017; Ganea and Hofmann, 2017) and often do not use virtually any other hand-crafted features. Ganea and Hofmann (2017) showed that such an approach can yield SOTA accuracy on a standard benchmark (AIDA-CoNLL dataset). Their local and pairwise score functions are

$$\Psi(e_i, c_i) = \mathbf{e}_i^T \mathbf{B} f(c_i)$$
$$\Phi(e_i, e_j, D) = \frac{1}{n-1} \mathbf{e}_i^T \mathbf{R} \mathbf{e}_j \qquad (3)$$

where $\mathbf{e}_i, \mathbf{e}_j \in \mathbb{R}^d$ are the embeddings of entity e_i, e_j, $\mathbf{B}, \mathbf{R} \in \mathbb{R}^{d \times d}$ are diagonal matrices. The mapping $f(c_i)$ applies an attention mechanism to context words in c_i to obtain a feature representations of context ($f(c_i) \in \mathbb{R}^d$).

Note that the global component (the pairwise scores) is agnostic to any relations between entities or even to their ordering: it models $e_1, ..., e_n$ simply as a bag of entities. Our work is in line with Ganea and Hofmann (2017) in the sense that feature engineering plays no role in computing local and pair-wise scores. Furthermore, we argue that pair-wise scores should take into account relations between mentions which are represented by *relation embeddings*.

3 Multi-relational models

3.1 General form

We assume that there are K latent relations. Each relation k is assigned to a mention pair (m_i, m_j) with a non-negative weight ('confidence') α_{ijk}. The pairwise score (m_i, m_j) is computed as a weighted sum of relation-specific pairwise scores (see Figure 2, top):

$$\Phi(e_i, e_j, D) = \sum_{k=1}^{K} \alpha_{ijk} \Phi_k(e_i, e_j, D)$$

$\Phi_k(e_i, e_j, D)$ can be any pairwise score function, but here we adopt the one from Equation 3. Namely, we represent each relation k by a diagonal matrix $\mathbf{R}_k \in \mathbb{R}^{d \times d}$, and

$$\Phi_k(e_i, e_j, D) = \mathbf{e}_i^T \mathbf{R}_k \mathbf{e}_j$$

The weights α_{ijk} are normalized scores:

$$\alpha_{ijk} = \frac{1}{Z_{ijk}} \exp\left\{\frac{f^T(m_i, c_i)\mathbf{D}_k f(m_j, c_j)}{\sqrt{d}}\right\} \quad (4)$$

where Z_{ijk} is a normalization factor, $f(m_i, c_i)$ is a function mapping (m_i, c_i) onto $\mathbb{R}^d$, and $\mathbf{D}_k \in \mathbb{R}^{d\times d}$ is a diagonal matrix.

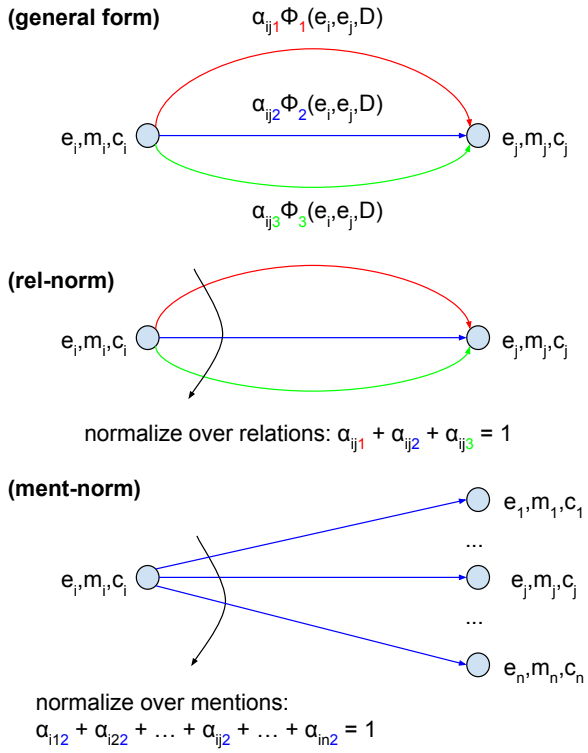

Figure 2: Multi-relational models: general form (top), rel-norm (middle) and ment-norm (bottom). Each color corresponds to one relation.

In our experiments, we use a single-layer neural network as f (see Figure 3) where c_i is a concatenation of the average embedding of words in the left context with the average embedding of words in the right context of the mention.[1]

As α_{ijk} is indexed both by mention index j and relation index k, we have two choices for Z_{ijk}: normalization over relations and normalization over mentions. We consider both versions of the model.

[1]We also experimented with LSTMs but we could not prevent them from severely overfitting, and the results were poor.

3.2 Rel-norm: Relation-wise normalization

For rel-norm, coefficients α_{ijk} are normalized over relations k, in other words,

$$Z_{ijk} = \sum_{k'=1}^{K} \exp\left\{\frac{f^T(m_i, c_i)\mathbf{D}_{k'} f(m_j, c_j)}{\sqrt{d}}\right\}$$

so that $\sum_{k=1}^{K} \alpha_{ijk} = 1$ (see Figure 2, middle). We can also re-write the pairwise scores as

$$\Phi(e_i, e_j, D) = \mathbf{e}_i^T \mathbf{R}_{ij} \mathbf{e}_j \quad (5)$$

where $\mathbf{R}_{ij} = \sum_{k=1}^{K} \alpha_{ijk}\mathbf{R}_k$.

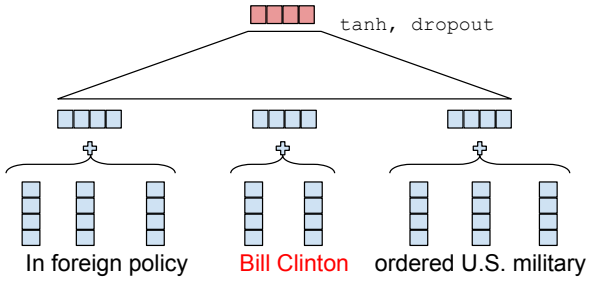

Figure 3: Function $f(m_i, c_i)$ is a single-layer neural network, with tanh activation function and a layer of dropout on top.

Intuitively, α_{ijk} is the probability of assigning a k-th relation to a mention pair (m_i, m_j). For every pair rel-norm uses these probabilities to choose one relation from the pool and relies on the corresponding relation embedding $\mathbf{R}_k$ to compute the compatibility score.

For $K = 1$ rel-norm reduces (up to a scaling factor) to the bag-of-entities model defined in Equation 3.

In principle, instead of relying on the linear combination of relation embeddings matrices $\mathbf{R}_k$, we could directly predict a context-specific relation embedding $\mathbf{R}_{ij} = diag\{g(m_i, c_i, m_j, c_j)\}$ where g is a neural network. However, in preliminary experiments we observed that this resulted in overfitting and poor performance. Instead, we choose to use a small fixed number of relations as a way to constrain the model and improve generalization.

3.3 Ment-norm: Mention-wise normalization

We can also normalize α_{ijk} over j:

$$Z_{ijk} = \sum_{\substack{j'=1 \\ j'\neq i}}^{n} \exp\left\{\frac{f^T(m_i, c_i)\mathbf{D}_k f(m_{j'}, c_{j'})}{\sqrt{d}}\right\}$$

This implies that $\sum_{j=1,j\neq i}^{n} \alpha_{ijk} = 1$ (see Figure 2, bottom). If we rewrite the pairwise scores as

$$\Phi(e_i, e_j, D) = \sum_{k=1}^{K} \alpha_{ijk} \mathbf{e}_i^T \mathbf{R}_k \mathbf{e}_j, \qquad (6)$$

we can see that Equation 3 is a special case of ment-norm when $K = 1$ and $\mathbf{D}_1 = \mathbf{0}$. In other words, Ganea and Hofmann (2017) is our mono-relational ment-norm with uniform α.

The intuition behind ment-norm is that for each relation k and mention m_i, we are looking for mentions related to m_i with relation k. For each pair of m_i and m_j we can distinguish two cases: (i) α_{ijk} is small for all k: m_i and m_j are not related under any relation, (ii) α_{ijk} is large for one or more k: there are one or more relations which are predicted for m_i and m_j.

In principle, rel-norm can also indirectly handle both these cases. For example, it can master (i) by dedicating a distinct 'none' relation to represent lack of relation between the two mentions (with the corresponding matrix $\mathbf{R}_k$ set to $\mathbf{0}$). Though it cannot assign large weights (i.e., close to 1) to multiple relations (as needed for (ii)), it can in principle use the 'none' relation to vary the probability mass assigned to the rest of relations across mention pairs, thus achieving the same effect (up to a multiplicative factor). Nevertheless, in contrast to ment-norm, we do not observe this behavior for rel-norm in our experiments: the inductive basis seems to disfavor such configurations.

Ment-norm is in line with the current trend of using the attention mechanism in deep learning (Bahdanau et al., 2014), and especially related to multi-head attention of Vaswani et al. (2017). For each mention m_i and for each k, we can interpret α_{ijk} as the probability of choosing a mention m_j among the set of mentions in the document. Because here we have K relations, each mention m_i will have maximally K mentions (i.e. heads in terminology of Vaswani et al. (2017)) to focus on. Note though that they use multi-head attention for choosing input features in each layer, whereas we rely on this mechanism to compute pairwise scoring functions for the structured output (i.e. to compute potential functions in the corresponding undirected graphical model, see Section 3.4).

Mention padding

A potentially serious drawback of ment-norm is that the model uses all K relations even in cases where some relations are inapplicable. For example, consider applying relation *coreference* to mention "West Germany" in Figure 1. The mention is non-anaphoric: there are no mentions co-referent with it. Still the ment-norm model has to distribute the weight across the mentions. This problem occurs because of the normalization $\sum_{j=1,j\neq i}^{n} \alpha_{ijk} = 1$. Note that this issue does not affect standard applications of attention: normally the attention-weighted signal is input to another transformation (e.g., a flexible neural model) which can then disregard this signal when it is useless. This is not possible within our model, as it simply uses α_{ijk} to weight the bilinear terms without any extra transformation.

Luckily, there is an easy way to circumvent this problem. We add to each document a padding mention m_{pad} linked to a padding entity e_{pad}. In this way, the model can use the padding mention to damp the probability mass that the other mentions receive. This method is similar to the way some mention-ranking coreference models deal with non-anaphoric mentions (e.g. Wiseman et al. (2015)).

3.4 Implementation

Following Ganea and Hofmann (2017) we use Equation 2 to define a conditional random field (CRF). We use the local score function identical to theirs and the pairwise scores are defined as explained above:

$$q(E|D) \propto \exp\left\{ \sum_{i=1}^{n} \Psi(e_i, c_i) + \sum_{i\neq j} \Phi(e_i, e_j, D) \right\}$$

We also use max-product loopy belief propagation (LBP) to estimate the max-marginal probability

$$\hat{q}_i(e_i|D) \approx \max_{\substack{e_1,\ldots,e_{i-1} \\ e_{i+1},\ldots,e_n}} q(E|D)$$

for each mention m_i. The final score function for m_i is given by:

$$\rho_i(e) = g(\hat{q}_i(e|D), \hat{p}(e|m_i))$$

where g is a two-layer neural network and $\hat{p}(e|m_i)$ is the probability of selecting e conditioned only on m_i. This probability is computed by mixing mention-entity hyperlink count statistics from Wikipedia, a large Web corpus and YAGO.[2]

[2] See Ganea and Hofmann (2017, Section 6).

We minimize the following ranking loss:

$$L(\theta) = \sum_{D \in \mathcal{D}} \sum_{m_i \in D} \sum_{e \in C_i} h(m_i, e) \quad (7)$$
$$h(m_i, e) = \max\left(0, \gamma - \rho_i(e_i^*) + \rho_i(e)\right)$$

where θ are the model parameters, $\mathcal{D}$ is a training dataset, and e_i^* is the ground-truth entity. Adam (Kingma and Ba, 2014) is used as an optimizer.

For ment-norm, the padding mention is treated like any other mentions. We add $\mathbf{f}_{pad} = f(m_{pad}, c_{pad})$ and $\mathbf{e}_{pad} \in \mathbb{R}^d$, an embedding of e_{pad}, to the model parameter list, and tune them while training the model.

In order to encourage the models to explore different relations, we add the following regularization term to the loss function in Equation 7:

$$\lambda_1 \sum_{i,j} \text{dist}(\mathbf{R}_i, \mathbf{R}_j) + \lambda_2 \sum_{i,j} \text{dist}(\mathbf{D}_i, \mathbf{D}_j)$$

where λ_1, λ_2 are set to -10^{-7} in our experiments, $\text{dist}(\mathbf{x}, \mathbf{y})$ can be any distance metric. We use:

$$\text{dist}(\mathbf{x}, \mathbf{y}) = \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_2} - \frac{\mathbf{y}}{\|\mathbf{y}\|_2} \right\|_2$$

Using this regularization to favor diversity is important as otherwise relations tend to collapse: their relation embeddings $\mathbf{R}_k$ end up being very similar to each other.

4 Experiments

We evaluated four models: (i) *rel-norm* proposed in Section 3.2; (ii) *ment-norm* proposed in Section 3.3; (iii) *ment-norm ($K = 1$)*: the mono-relational version of ment-norm; and (iv) *ment-norm (no pad)*: the ment-norm without using mention padding. Recall also that our mono-relational (i.e. $K = 1$) rel-norm is equivalent to the relation-agnostic baseline of Ganea and Hofmann (2017).

We implemented our models in PyTorch and run experiments on a Titan X GPU. The source code and trained models will be publicly available at https://github.com/lephong/mulrel-nel.

4.1 Setup

We set up our experiments similarly to those of Ganea and Hofmann (2017), run each model 5 times, and report average and 95% confidence interval of the standard micro F1 score (aggregates over all mentions).

Datasets

For *in-domain* scenario, we used AIDA-CoNLL dataset[3] (Hoffart et al., 2011). This dataset contains AIDA-train for training, AIDA-A for dev, and AIDA-B for testing, having respectively 946, 216, and 231 documents. For *out-domain* scenario, we evaluated the models trained on AIDA-train, on five popular test sets: MSNBC, AQUAINT, ACE2004, which were cleaned and updated by Guo and Barbosa (2016); WNED-CWEB (CWEB), WNED-WIKI (WIKI), which were automatically extracted from ClueWeb and Wikipedia (Guo and Barbosa, 2016; Gabrilovich et al., 2013). The first three are small with 20, 50, and 36 documents whereas the last two are much larger with 320 documents each. Following previous works (Yamada et al., 2016; Ganea and Hofmann, 2017), we considered only mentions that have entities in the KB (i.e., Wikipedia).

Candidate selection

For each mention m_i, we selected 30 top candidates using $\hat{p}(e|m_i)$. We then kept 4 candidates with the highest $\hat{p}(e|m_i)$ and 3 candidates with the highest scores $\mathbf{e}^T \left(\sum_{w \in d_i} \mathbf{w}\right)$, where $\mathbf{e}, \mathbf{w} \in \mathbb{R}^d$ are entity and word embeddings, d_i is the 50-word window context around m_i.

Hyper-parameter setting

We set $d = 300$ and used GloVe (Pennington et al., 2014) word embeddings trained on 840B tokens for computing f in Equation 4, and entity embeddings from Ganea and Hofmann (2017).[4] We use the following parameter values: $\gamma = 0.01$ (see Equation 7), the number of LBP loops is 10, the dropout rate for f was set to 0.3, the window size of local contexts c_i (for the pairwise score functions) is 6. For rel-norm, we initialized $\text{diag}(\mathbf{R}_k)$ and $\text{diag}(\mathbf{D}_k)$ by sampling from $\mathcal{N}(0, 0.1)$ for all k. For ment-norm, we did the same except that $\text{diag}(\mathbf{R}_1)$ was sampled from $\mathcal{N}(1, 0.1)$.

To select the best number of relations K, we considered all values of $K \leq 7$ ($K > 7$ would not fit in our GPU memory, as some of the documents are large). We selected the best ones based on the development scores: 6 for rel-norm, 3 for ment-norm, and 3 for ment-norm (no pad).

When training the models, we applied early stopping. For rel-norm, when the model reached

[3] TAC KBP datasets are no longer available.

[4] https://github.com/dalab/deep-ed

91% F1 on the dev set, [5] we reduced the learning rate from 10^{-4} to 10^{-5}. We then stopped the training when F1 was not improved after 20 epochs. We did the same for ment-norm except that the learning rate was changed at 91.5% F1.

Note that all the hyper-parameters except K and the turning point for early stopping were set to the values used by Ganea and Hofmann (2017). Systematic tuning is expensive though may have further increased the result of our models.

4.2 Results

Methods	Aida-B
Chisholm and Hachey (2015)	88.7
Guo and Barbosa (2016)	89.0
Globerson et al. (2016)	91.0
Yamada et al. (2016)	91.5
Ganea and Hofmann (2017)	92.22 ± 0.14
rel-norm	92.41 ± 0.19
ment-norm	$\mathbf{93.07} \pm 0.27$
ment-norm ($K = 1$)	92.89 ± 0.21
ment-norm (no pad)	92.37 ± 0.26

Table 1: F1 scores on AIDA-B (test set).

Table 1 shows micro F1 scores on AIDA-B of the SOTA methods and ours, which all use Wikipedia and YAGO mention-entity index. To our knowledge, ours are the only (unsupervisedly) inducing and employing more than one relations on this dataset. The others use only one relation, coreference, which is given by simple heuristics or supervised third-party resolvers. All four our models outperform any previous method, with ment-norm achieving the best results, 0.85% higher than that of Ganea and Hofmann (2017).

Table 2 shows micro F1 scores on 5 out-domain test sets. Besides ours, only Cheng and Roth (2013) employs several mention relations. Ment-norm achieves the highest F1 scores on MSNBC and ACE2004. On average, ment-norm's F1 score is 0.3% higher than that of Ganea and Hofmann (2017), but 0.2% lower than Guo and Barbosa (2016)'s. It is worth noting that Guo and Barbosa (2016) performs exceptionally well on WIKI, but substantially worse than ment-norm on all other datasets. Our other three models, however, have lower average F1 scores compared to the best previous model.

The experimental results show that ment-norm outperforms rel-norm, and that mention padding plays an important role.

[5] We chose the highest F1 that rel-norm always achieved without the learning rate reduction.

4.3 Analysis

Mono-relational v.s. multi-relational

For rel-norm, the mono-relational version (i.e., Ganea and Hofmann (2017)) is outperformed by the multi-relational one on AIDA-CoNLL, but performs significantly better on all five out-domain datasets. This implies that multi-relational rel-norm does not generalize well across domains.

For ment-norm, the mono-relational version performs worse than the multi-relational one on all test sets except AQUAINT. We speculate that this is due to multi-relational ment-norm being less sensitive to prediction errors. Since it can rely on multiple factors more easily, a single mistake in assignment is unlikely to have large influence on its predictions.

Oracle

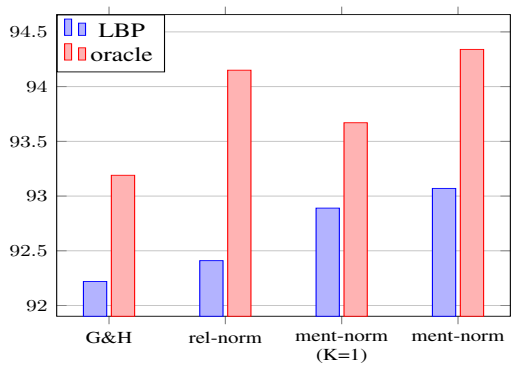

Figure 4: F1 on AIDA-B when using LBP and the oracle. G&H is Ganea and Hofmann (2017).

In order to examine learned relations in a more transparant setting, we consider an idealistic scenario where imperfection of LBP, as well as mistakes in predicting other entities, are taken out of the equation using an oracle. This oracle, when we make a prediction for mention m_i, will tell us the correct entity e_j^* for every other mentions $m_j, j \neq i$. We also used AIDA-A (development set) for selecting the numbers of relations for rel-norm and ment-norm. They are set to 6 and 3, respectively. Figure 4 shows the micro F1 scores.

Surprisingly, the performance of oracle rel-norm is close to that of oracle ment-norm, although without using the oracle the difference was substantial. This suggests that rel-norm is more sensitive to prediction errors than ment-norm. Ganea and Hofmann (2017), even with the help of the oracle, can only perform slightly better than LBP (i.e. non-oracle) ment-norm. This

Methods	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	Avg
Milne and Witten (2008)	78	85	81	64.1	81.7	77.96
Hoffart et al. (2011)	79	56	80	58.6	63	67.32
Ratinov et al. (2011)	75	83	82	56.2	67.2	72.68
Cheng and Roth (2013)	90	**90**	86	67.5	73.4	81.38
Guo and Barbosa (2016)	92	87	88	77	**84.5**	**85.7**
Ganea and Hofmann (2017)	93.7 ± 0.1	88.5 ± 0.4	88.5 ± 0.3	**77.9** ± 0.1	77.5 ± 0.1	85.22
rel-norm	92.2 ± 0.3	86.7 ± 0.7	87.9 ± 0.3	75.2 ± 0.5	76.4 ± 0.3	83.67
ment-norm	**93.9** ± 0.2	88.3 ± 0.6	**89.9** ± 0.8	77.5 ± 0.1	$\underline{78.0} \pm 0.1$	$\underline{85.51}$
ment-norm ($K = 1$)	93.2 ± 0.3	88.4 ± 0.4	88.9 ± 1.0	77.0 ± 0.2	77.2 ± 0.1	84.94
ment-norm (no pad)	93.6 ± 0.3	87.8 ± 0.5	**90.0** ± 0.3	77.0 ± 0.2	77.3 ± 0.3	85.13

Table 2: F1 scores on five out-domain test sets. Underlined scores show cases where the corresponding model outperforms the baseline.

suggests that its global coherence scoring component is indeed too simplistic. Also note that both multi-relational oracle models substantially outperform the two mono-relational oracle models. This shows the benefit of using more than one relations, and the potential of achieving higher accuracy with more accurate inference methods.

Relations

In this section we qualitatively examine relations that the models learned by looking at the probabilities α_{ijk}. See Figure 5 for an example. In that example we focus on mention "Liege" in the sentence at the top and study which mentions are related to it under two versions of our model: rel-norm (leftmost column) and ment-norm (rightmost column).

For rel-norm it is difficult to interpret the meaning of the relations. It seems that the first relation dominates the other two, with very high weights for most of the mentions. Nevertheless, the fact that rel-norm outperforms the baseline suggests that those learned relations encode some useful information.

For ment-norm, the first relation is similar to coreference: the relation prefers those mentions that potentially refer to the same entity (and/or have semantically similar mentions): see Figure 5 (left, third column). The second and third relations behave differently from the first relation as they prefer mentions having more distant meanings and are complementary to the first relation. They assign large weights to (1) "Belgium" and (2) "Brussels" but small weights to (4) and (6) "Liege". The two relations look similar in this example, however they are not identical in general. See a histogram of bucketed values of their weights in Figure 5 (right): their α have quite different distributions.

Complexity

The complexity of rel-norm and ment-norm is linear in K, so in principle our models should be considerably more expensive than Ganea and Hofmann (2017). However, our models converge much faster than their relation-agnostic model: on average ours needs 120 epochs, compared to theirs 1250 epochs. We believe that the structural bias helps the model to capture necessary regularities more easily. In terms of wall-clock time, our model requires just under 1.5 hours to train, that is ten times faster than the relation agnostic model (Ganea and Hofmann, 2017). In addition, the difference in testing time is negligible when using a GPU.

5 Conclusion and Future work

We have shown the benefits of using relations in NEL. Our models consider relations as latent variables, thus do not require any extra supervision. Representation learning was used to learn relation embeddings, eliminating the need for extensive feature engineering. The experimental results show that our best model achieves the best reported F1 on AIDA-CoNLL with an improvement of 0.85% F1 over the best previous results.

Conceptually, modeling multiple relations is substantially different from simply modeling coherence (as in Ganea and Hofmann (2017)). In this way we also hope it will lead to interesting follow-up work, as individual relations can be informed by injecting prior knowledge (e.g., by training jointly with relation extraction models).

In future work, we would like to use syntactic and discourse structures (e.g., syntactic dependency paths between mentions) to encourage the models to discover a richer set of relations. We also would like to combine ment-norm and rel-norm. Besides, we would like to examine whether

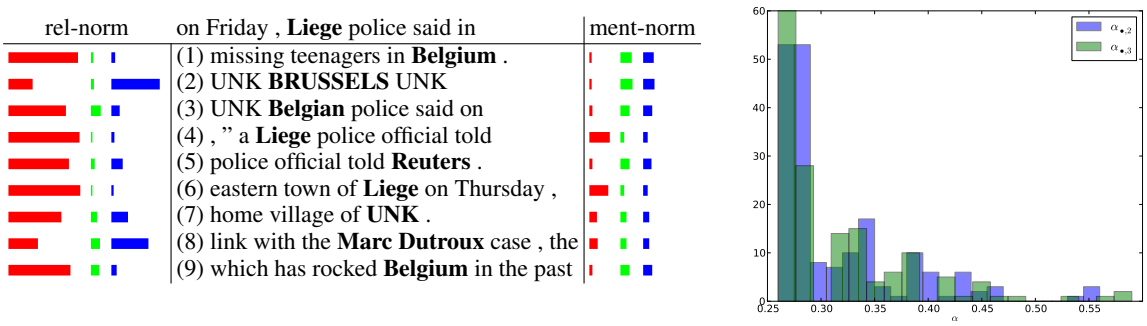

Figure 5: (Left) Examples of α. The first and third columns show α_{ijk} for oracle rel-norm and oracle ment-norm, respectively. (Right) Histograms of $\alpha_{\bullet k}$ for $k = 2, 3$, corresponding to the second and third relations from oracle ment-norm. Only $\alpha > 0.25$ (i.e. high attentions) are shown.

the induced latent relations could be helpful for relation extract.

Acknowledgments

We would like to thank anonymous reviewers for their suggestions and comments. The project was supported by the European Research Council (ERC StG BroadSem 678254), the Dutch National Science Foundation (NWO VIDI 639.022.518), and an Amazon Web Services (AWS) grant.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, Seattle, Washington, USA. Association for Computational Linguistics.

Andrew Chisholm and Ben Hachey. 2015. Entity disambiguation with web links. *Transactions of the Association of Computational Linguistics*, 3:145–156.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of clueweb corpora.

Octavian-Eugen Ganea, Marina Ganea, Aurelien Lucchi, Carsten Eickhoff, and Thomas Hofmann. 2016. Probabilistic bag-of-hyperlinks model for entity linking. In *Proceedings of the 25th International Conference on World Wide Web*, pages 927–938. International World Wide Web Conferences Steering Committee.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2609–2619. Association for Computational Linguistics.

Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringaard, and Fernando Pereira. 2016. Collective entity resolution with multi-focal attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631. Association for Computational Linguistics.

Zhaochen Guo and Denilson Barbosa. 2016. Robust named entity disambiguation with random walks. *Semantic Web*, (Preprint).

Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–34, Sofia, Bulgaria. Association for Computational Linguistics.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011.

Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.

Hongzhao Huang, Larry Heck, and Heng Ji. 2015. Leveraging deep neural networks and knowledge graphs for entity disambiguation. *arXiv preprint arXiv:1504.07678*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Nevena Lazic, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. 2015. Plato: A selective context model for entity resolution. *Transactions of the Association for Computational Linguistics*, 3:503–515.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

Kevin P Murphy, Yair Weiss, and Michael I Jordan. 1999. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384. Association for Computational Linguistics.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1015–1024. International World Wide Web Conferences Steering Committee.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Martin J Wainwright, Michael I Jordan, et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426. Association for Computational Linguistics.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259. Association for Computational Linguistics.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2017. Learning distributed representations of texts and entities from knowledge base. *arXiv preprint arXiv:1705.02494*.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1321–1331, Beijing, China. Association for Computational Linguistics.